Principles and Practice of Explainable Machine Learning

Year: 2020 | Citations: 516 | Authors: Vaishak Belle, I. Papantonis

Abstract

Artificial intelligence (AI) provides many opportunities to improve private and public life. Discovering patterns and structures in large troves of data in an automated manner is a core component of data science, and currently drives applications in diverse areas such as computational biology, law and finance. However, such a highly positive impact is coupled with a significant challenge: how do we understand the decisions suggested by these systems in order that we can trust them? In this report, we focus specifically on data-driven methods—machine learning (ML) and pattern recognition models in particular—so as to survey and distill the results and observations from the literature. The purpose of this report can be especially appreciated by noting that ML models are increasingly deployed in a wide range of businesses. However, with the increasing prevalence and complexity of methods, business stakeholders in the very least have a growing number of concerns about the drawbacks of models, data-specific biases, and so on. Analogously, data science practitioners are often not aware about approaches emerging from the academic literature or may struggle to appreciate the differences between different methods, so end up using industry standards such as SHAP. Here, we have undertaken a survey to help industry practitioners (but also data scientists more broadly) understand the field of explainable machine learning better and apply the right tools. Our latter sections build a narrative around a putative data scientist, and discuss how she might go about explaining her models by asking the right questions. From an organization viewpoint, after motivating the area broadly, we discuss the main developments, including the principles that allow us to study transparent models vs. opaque models, as well as model-specific or model-agnostic post-hoc explainability approaches. We also briefly reflect on deep learning models, and conclude with a discussion about future research directions.